Length Polymorphisms of Simple Sequence Repeat DNA in Soybean

Mahinur S. Akkaya, Arvind A. Bhagwat[1] and Perry B. Cregan[2]

Soybean and Alfalfa Research Laboratory, United States Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center, Beltsville, Maryland 20705-2350

Manuscript received May 14, 1992
Accepted for publication August 8, 1992

ABSTRACT

The objective of this work was to ascertain the presence and degree of simple sequence repeat (SSR) DNA length polymorphism in the soybean [*Glycine max* (L.) Merr.]. A search of GenBank revealed no $(CA)_n$ or $(GT)_n$ SSRs with n greater than 8 in soybean. In contrast, 5 $(AT)_n$ and 1 $(ATT)_n$ SSRs with n ranging from 14 to 27 were detected. Polymerase chain reaction (PCR) primers to regions flanking the six SSR loci were used in PCR amplification of DNA from 43 homozygous soybean genotypes. At three loci, amplification produced one PCR product per genotype and revealed 6, 7 and 8 product length variants (alleles) at the three loci, respectively. F_1 hybrids between parents carrying different alleles produced two PCR products identical to the two parents. Codominant segregation of alleles among F_2 progeny was demonstrated at each locus. A soybean DNA library was screened for the presence of $(CA/GT)_n$ SSRs. Sequencing of positive clones revealed that the longest such SSR was $(CA)_9$. Thus, $(CA)_n$ SSRs with n of 15 or more are apparently much less common in soybean than in the human genome. In contrast to humans, $(CA)_n$ SSRs will probably not provide an abundant source of genetic markers in soybean. However, the apparent abundance of long $(AT)_n$ sequences should allow this SSR to serve as a source of highly polymorphic genetic markers in soybean.

REPETITIVE DNA sequences such as variable number tandem repeat (VNTR) loci (NAKAMURA *et al.* 1987) serve as highly informative genetic markers. These multiallelic loci consist of repeated core sequences which have been referred to as minisatellites (JEFFERYS, WILSON and THEIN 1985). The tandemly repeated minisatellites are flanked by conserved endonuclease restriction sites. Thus, the length of the restriction fragment produced by this type of genetic locus is proportional to the number of core units it contains. It was subsequently suggested (JEFFREYS *et al.* 1988) that the highly informative nature of VNTR loci be combined with the specificity and rapidity of polymerase chain reaction (PCR) technology (MULLIS *et al.* 1986). Primers to the conserved flanking regions of VNTR loci were developed allowing amplification of the entire VNTR locus. Resulting PCR products possess electrophoretic mobilities which differ according to the number of core units in the VNTR allele(s) present.

This approach was recently extended to a different type of repetitive DNA in humans (LITT and LUTY 1989; WEBER and MAY 1989; TAUTZ 1989). Rather than repeat units in the range of 11–60 bp in length as occur in the minisatellites, these workers suggested that high levels of polymorphism exist in dinucleotide tandem repeat sequences. For example, $(CA/GT)_n$ were reported to occur in the human genome as many as 50,000 times with n varying from 10 to 60. This type of reiterated sequence has been termed a simple sequence repeat (SSR) by JACOB *et al.* (1991), short tandem repeat by EDWARDS *et al.* (1991), or microsatellite (LITT and LUTY 1989). The DNA sequences flanking SSRs are conserved, allowing the selection of PCR primers that will amplify the intervening SSR. As initially reported (LITT and LUTY 1989; WEBER and MAY 1989; TAUTZ 1989), the PCR reaction includes one ^{32}P-labeled nucleotide or one or two ^{32}P end-labeled primers to allow visualization of amplification products via autoradiography after electrophoresis on a standard sequencing gel. Variation in PCR product length is a function of the number of SSR units. SSR length polymorphism is detected when the PCR products from a particular locus differ in length. This type of genetic marker can be placed on genetic maps in relation to other SSR, restriction fragment length polymorphism (RFLP) (BOTSTEIN *et al.* 1980), random amplified polymorphic DNA (RAPD) (WILLIAMS *et al.* 1990), and phenotypic markers in a manner identical to that used with RFLP, RAPD, and phenotypic markers.

Human geneticists first demonstrated the highly polymorphic nature of SSRs in 1989 (LITT and LUTY 1989; WEBER and MAY 1989), with up to 12 alleles at some SSR loci. SSRs include tri- and tetrameric repeats such as $(AAT)_n$ and $(AGAT)_n$. According to EDWARDS *et al.* (1991), the combined frequency in the human genome of tri- and tetrameric SSRs with n = 7 or greater is estimated to be 400,000 or one such SSR per 10 kbp.

[1] Current address: Agronomy Department, University of Maryland, College Park, Maryland 20742.
[2] To whom reprint requests should be sent.

The presence and abundance of SSR loci in higher plants is not well defined. A search of only 35 kbp of algal and plant DNA sequence data (TAUTZ, TRICK and DOVER 1986) indicated similar numbers of di- and trinucleotide repeats per 100 kbp as occurs in vertebrates. The only report documenting SSRs in higher plants is that of CONDIT and HUBBELL (1991). They screened DNA libraries of five tropical tree species as well as *Zea mays* for the presence of clones containing $(AC)_n$ and $(AG)_n$ sequences and estimated a total of between 5×10^3 and 3×10^5 such sequences in the species examined. We are not aware of any published reports in plants regarding length polymorphism of SSRs.

Based upon the abundance and informativeness of SSR markers in humans and the lack of information concerning the frequency and polymorphic nature of SSRs in plants and particularly in soybean [*Glycine max* (L.) Merr.], we undertook the study presented here. The experiments reported here were designed to investigate: (1) the frequency and nature of SSRs in soybean as determined by a search of GenBank; (2) the presence of SSR length polymorphism in soybean and wild soybean (*Glycine soja* Sieb. & Zucc.); (3) the mode of inheritance of SSR markers in soybean; and (4) if the frequency and length of $(CA/GT)_n$ SSRs in soybean is comparable to that found in the human genome.

MATERIALS AND METHODS

Search of GenBank soybean sequences for SSRs: At the time of our search, GenBank had a total of 141 soybean sequences. Of these, 10 appeared to be chloroplast-derived and at least one was a duplicate. The remaining 130 sequences were searched for all possible di-, tri- and tetrameric SSRs. In the case of dinucleotide repeats, only those with greater than four repeat units were selected. For tri- and tetrameric SSRs, we selected those with greater than three repeat units.

Soybean DNA library, colony hybridization and probing: Soybean DNA (cv. Williams) fragments of 250–300 bp in length obtained by digestion with restriction enzymes *Xba*I and *Hin*dIII (U.S. Biochemical Corp., Cleveland, Ohio) were cloned into pBluescript II KS+ (STRATAGENE, LaJolla, California) which were used to transform *Escherichia coli* host strain XL1-Blue followed by blue/white color selection. Selected colonies were picked into microtiter plates containing LB medium with ampicillin (50 μg/ml) and tetracycline (10 μg/ml). The library was screened by colony hybridization. Procedures outlined by SAMBROOK, FRITSCH and MANIATIS (1989) were followed throughout. Filters were hybridized to a ^{32}P-labeled $G(AC)_{11}$ oligonucleotide probe which was labeled using a reaction mix containing a $(GT)_{11}C$ oligonucleotide template, a $G(AC)_3$ primer, Klenow fragment, and [α-^{32}P]dATP and [α-^{32}P]dCTP and filtered through a BioSpin 6 chromatography column (Rio-Rad, Richmond, California) for removal of unincorporated nucleotides. Hybridization conditions were as described by ROSTAS *et al.* (1986). Membranes were washed for 1–1.5 hr at 42° in 2 × SSC and exposed to x-ray film.

The selection and synthesis of PCR primers: WEBER (1990) indicated that in humans dinucleotide sequences with 10 or fewer repeats were unlikely to be polymorphic. Therefore, we synthesized primers to regions flanking only those GenBank dinucleotide SSRs with greater than 10 repeats. Likewise, EDWARDS *et al.* (1991) indicated that in the human populations they investigated approximately seven tri- or tetrameric repeats were necessary to produce a 50% probability of polymorphism. Therefore, primers were only synthesized when the number of tri- or tetranucleotide repeat units met these minimum length requirements. In the case of SSRs obtained from the Bluescript library no such minimum length requirement was imposed.

To select PCR primers to regions flanking SSRs, sequence data were analyzed using Primer Detective, Primer Design Software Program (CLONTECH Laboratories, Inc., Palo Alto, California). Primers were selected with expected PCR product length between 130 and 300 bp and were synthesized on an Applied Biosystems International 391 DNA Synthesizer followed by high performance liquid chromatography purification. In addition to the PCR primers the oligonucleotides $(GT)_{11}C$ and $G(AC)_8$ were synthesized for use in the development of the radiolabeled $G(AC)_{11}$ probe.

Soybean genotypes, soybean hybridization and DNA isolation: A group of 43 soybean genotypes (Table 1) were selected to use for investigating the presence of length polymorphism of SSRs using the primer sets that were synthesized (see Table 2). These soybeans can be classified into five groups: (1) northern United States ancestral cultivars that include genotypes that formed the germplasm base of cultivars grown in the northern United States and Canada, (2) northern U.S. cultivars that include soybeans developed via hybridization from the northern U.S. ancestral cultivars through a number of breeding cycles, (3) southern U.S. ancestral cultivars that include genotypes that form the germplasm base of cultivars grown in the southern United States, (4) southern U.S. cultivars that include soybeans developed via hybridization from the southern U.S. ancestral cultivars through a number of breeding cycles, and (5) *G. soja* (wild soybean) from a range of origins in the Far East. Seeds were originally obtained from Dr. Randall Nelson, USDA-ARS, University of Illinois, Department of Agronomy, 1102 South Goodwin Avenue, Urbana, Illinois 61801.

The 10 possible crosses between the soybean cultivars Amsoy, Williams, Fiskeby V, Tokyo and Jackson were made in the field or in the photoperiod houses at Beltsville, Maryland (BORTHWICH and PARKER 1952). F_1 plants of each cross were grown in the greenhouse for DNA isolation and to obtain F_2 seeds. Approximately 70 F_2 plants of the cross Williams × Amsoy and 100 plants of Jackson × Williams, along with the 43 soybean genotypes described earlier were grown in the greenhouse. Plants were inoculated with *Bradyrhizobium japonicum* strain USDA 110. DNA was isolated from leaf tissue of these plants using the procedure described by SAGHAI-MAROOF *et al.* (1984).

Southern hybridization: Human placental and *E. coli* strain B DNA were obtained from the Sigma Chemical Co. (St. Louis, Missouri). Soybean (cv. Williams) DNA was isolated as described above. *Pal*I (Stratagene)-digested DNAs (3 μg/lane) were separated on a 0.7% agarose gel and transferred to a Nytran membrane (Schleicher & Schuell, Inc., Keene, New Hampshire) as described by (SAMBROOK, FRITSCH and MANIATIS 1989). After baking at 80° for 2 hr the membrane was prehybridized and hybridized to a ^{32}P-labeled $G(AC)_{11}$ probe as described above. The membrane was washed for 1 hr at 42° with 2 × SSC and exposed to x-ray film. The blot was further washed at increasingly stringent conditions including 1 hr at 52° in 2 × SSC, 1 hr at 62° in 2 × SSC, 1 hr at 65° in 1 × SSC, and 2 hr at 66° in

TABLE 1

Soybean cultivars and wild soybean accessions examined for the presence of simple sequence repeat length polymorphism

Northern U.S. ancestral cultivars		Northern U.S. cultivars				Southern U.S. ancestral cultivars		Southern U.S. cultivars		*G. soja* plant introductions	
Lane No.[a]	Soybean genotype	Lane No.[a]	Soybean genotype	Lane No.[a]	Soybean genotype	Lane No.[a]	Soybean genotype	Lane No.[a]	Soybean genotype	Lane No.[a]	Soybean genotype
1	Mandarin (Ottowa)	8	Lincoln	15	Amsoy	22	CNS	28	Ogden	39	PI 468397
2	Mandarin	9	Capital	16	Calland	23	S-100	29	Lee	40	PI 101404B
3	Manchu (Madison)	10	Adams	17	Williams	24	Roanoke	30	Hill	41	PI 378693B
4	Richland	11	Hawkeye	18	Harcor	25	Tokyo	31	Jackson	42	PI 342619A
5	A.K. (Harrow)	12	Harosoy	19	Douglas	26	PI 54610-4	32	Pickett	43	PI 349647
6	Mukden	13	Merit	20	Century	27	Palmetto	33	Dyer		
7	Dunfield	14	Clark	21	Fiskeby V			34	Bragg		
								35	Tracy		
								36	Forrest		
								37	Essex		
								38	Centennial		

[a] Lane No. refers to the lane number occupied by each genotype in Figure 1.

0.1 × SSC. The blot was exposed to x-ray film after each wash.

PCR: PCR (10 μl) contained 1 × PCR buffer (10 mM Tris-HCl, pH 8.3, 50 mM KCl), 1 mM Mg^{2+}, 0.1 mM of each dNTP, 1.5 pmol each of the oligonucleotide primers, 0.1 μl of 3000 Ci/mmol [α-^{32}P]dATP (Amersham Corp., Arlington Heights, Illinois), 1 unit of *Taq* DNA polymerase (U.S. Biochemical Corp.) and 20 ng of soybean DNA template. A two-step PCR protocol was employed with 30 cycles; denaturation (1 min, 94°) and annealing/extension (1.5 min, 62°) using a Perkin-Elmer Cetus thermocycler.

DNA sequencing and polyacrylamide gel electrophoresis: DNA sequencing reactions of selected positive clones from the colony hybridization were performed on alkaline denatured dsDNA, prepared as in KRAFT *et al.* (1988) using U.S. Biochemical Corp. DNA sequencing kit Version 2.0. Reaction products were separated on a denaturing gel containing 6% polyacrylamide, 8 M urea and 1 × TBE at 60-watt constant power using a DNA sequencing gel electrophoresis apparatus (GIBCO/BRL, Gaithersburg, Maryland). The PCR products (1–1.8 μl/lane) were separated the same way after addition of stop solution supplied with the sequencing kit and denaturation for 5–10 min at 95°. Gels were dried and exposed to a Kodak XAR-5 (Eastman Kodak, Rochester, New York) film for 1–2 hr. A and G sequencing reactions of M13 mp18 ssDNA were used as molecular weight standards (U.S. Biochemical Corp.).

RESULTS

The search of GenBank for SSR sequences in soybean identified a total of 33 sequences with at least five repeat units. Only four of these contained (CA/GT)$_n$ repeats and five contained (TC/AG)$_n$ SSRs with n ranging from 5 to 8. In contrast, 17 sequences had SSRs composed of (AT/TA)$_n$ repeat units with n ranging from 5 to 27. Six trimeric and one tetrameric repeat were also identified. Of the 33 SSRs only two were located within coding sequences and both of these were trimeric repeats. Primers flanking each of the six DNA repeat regions with greater than 10 dimeric or greater than 6 trimeric SSRs were synthesized (Table 2) and used in PCR amplification of DNA from the four soybean cultivars Amsoy, Williams, Tokyo and Fiskeby V.

Length polymorphism of SSRs: In the case of the SOYHSP176(AT)$_n$ (NAGAO *et al.* 1985), SOYSC514(AT)$_n$ (SHIBATA, KATO and TANAKA 1991), and SOYPRP1(ATT)$_n$ (HONG, NAGAO and KEY 1987) loci, PCR amplification with the four initial cultivars produced variable length products. Subsequent investigation of 38 cultivars and 5 *G. soja* genotypes revealed 6, 8 and 7 alleles at the three loci, respectively (Figure 1). The most common allele at each locus was arbitrarily assigned a length of Z nucleotides (nt). Other alleles were assigned values of $Z \pm$ nt with nt equaling the apparent number of nt greater or less than Z in the denatured PCR product (Table 3). The number of nucleotides in each allele was established using the size of the most intense band of each amplification product. The precise size of particular alleles could not always initially be determined with certainty. When uncertainty existed, gels were exposed to film for shorter or longer periods or PCR products were reanalyzed with the sequencing ladder in an adjacent lane. In the case of the dinucleotide repeat loci SOYHSP176(AT)$_n$ and SOYSC514(AT)$_n$, the difference between alleles varied by multiples of 2 while at the SOYPRP1(ATT)$_n$ locus, which contained a trimeric repeat, alleles varied by multiples of three (Table 3). A total of 21 alleles were detected at the three loci. Nine of these occurred only once among the 43 genotypes. Of these nine, eight were found in one of the five *G. soja* genotypes.

In the case of the SOYHSP176(AT)$_n$ locus, the most common allele had an estimated length of 180 nt and the predicted length from the GenBank sequence was 191 nt. The second most common allele at the locus had an apparent length of 192 nt (Z + 12). The SOYHSP176 sequence was determined from a clone of the cultivar Corsoy (NAGAO *et al.* 1985). The par-

TABLE 2

Soybean sequences from GenBank found to possess SSRs with greater than 10 repeat units in dinucleotide SSRs and greater than six in trinucleotide SSRs and sequences from the pBluescript library with $(CA/GT)_{>6}$, bases contained in each repeat unit, number of repeat units (n) in GenBank or Bluescript clone sequence, primers used in attempted PCR amplification, and predicted PCR product size based on sequence information

Locus and SSR sequence	No. of repeat units (n)	Sense primer	Antisense primer	Predicted PCR product size (nt)
GenBank locus				
Dinucleotide SSRs				
SOYABAB$(AT)_n$	24	gataacaaacataaaaaagg	cgggaatgattgcacattttgcagg	290
SOYGLO2$(AT)_n$	27	atacagcgttgtgttctacaattcg	gattagtggggtctactccatttgc	264
SOYGY2$(AT)_n$	20	tagatacagatagataaataagtaa	aataaattagagcaaatggtttggg	244
SOYHSP176$(AT)_n$	15	ttttgttaagttactgtactgtgg	tattttagcagttttagatgattcg	191
SOYSC514$(AT)_n$	14	ctacatgacacaattcttagggacc	tggaaatcagtggaatatgtgaagc	173
Trimeric SSRs				
SOYPRP1$(ATT)_n$	19	aagaggtacgtgccaaattacatca	atcttgttagaaaactccgcccaca	187
$(CA/GT)_n$ from pBluescript Library				
5A2-$(AC)_n$	7	gcttatgttgtaaggtgaaggc	taatcttgttatggagttcagg	133
16C1-$(CA)_n$	9	tctgtttccttctcagacaagctcc	cacaagaacactgctttctcctggc	216
45C5-$(AC)_n(AT)_n(N)_{90}(ATT)_n$	8, 4, 7	tctttggagccattcagacaaaagga	tagcttgatttgccagaaataaa	243
105H7-$(AC)_n$	7	aaacttttttgaaaacccc	agtagtattatgcaaagaatatccc	195

ents of Corsoy are the cultivars Harosoy and Capital which both carry the Z + 12 allele (Figure 1A, lanes 12 and 9, respectively). At the SOYSC514$(AT)_n$ locus, the most common allele had a length of 174 nt which was one nucleotide longer than that predicted based upon the GenBank sequence. In the case of SOYPRP1$(ATT)_n$ the most common allele was 160 nt in length whereas the predicted length from the GenBank sequence was 187 nt. The size of this allele was identical to the Z + 27 SOYPRP1$(ATT)_n$ allele which occurred in 7 of the 43 genotypes examined.

In the case of the SOYABAB$(AT)_n$ and SOYGY2$(AT)_n$ loci, PCR amplifications were unsuccessful with the DNA of any of the initial four soybeans as template. Using primers to the SOYGL02$(AT)_n$ locus, two PCR products were produced with each soybean genotype, one of which appeared to be polymorphic. Further study of these loci is being undertaken.

Mendelian inheritance of SSR alleles: In no case did any of the 43 soybean genotypes appear to be heterozygous, *i.e.*, produce more than one PCR product. Not surprisingly, a contrasting result was obtained with PCR amplification of DNA from F_1 plants of crosses between genotypes carrying different SSR alleles (Figure 2). In the 10 possible F_1 hybrids between cultivars Amsoy, Williams, Fiskeby V, Tokyo and Jackson, the F_1s always produced two PCR products in those cases in which the parents were polymorphic. Codominant segregation of SSR markers among F_2 soybean progeny was demonstrated at the three SSR loci. For example, at the SOYPRP1$(ATT)_n$ locus, Jackson carries the Z allele and Williams the Z + 27 allele. Of 98 F_2 progeny, 22 were homozygotes carrying the Z + 27 allele, 53 were heterozygotes, and 23 were homozygotes carrying the Z allele. This very closely fits the 1:2:1 ratio of codominant segregation with a Chi-squared value of 0.7 (P = 0.9–0.7). The allelic constitution of 38 of these F_2 progeny is illustrated in Figure 3. Similar codominant segregation ratios were obtained among the F_2 progeny of Jackson × Williams at the SOYSC514$(AT)_n$ locus and at all three loci in the cross of Williams × Amsoy (Table 4).

The presence of $(CA/GT)_n$ sequences in soybean: The first reports of SSR length polymorphism in humans were of $(CA/GT)_n$ sequences and therefore, concurrent with the search of Genbank for SSRs, we probed soybean DNA with a radiolabeled poly(CA) probe to determine if such sequences were also common in soybean. The poly(CA) probe hybridized to *PalI* digested human DNA and to a lesser extent to Williams soybean DNA (Figure 4), while no hybridization to *E. coli* was observed. The probe remained annealed to human and soybean DNA even after very high stringency washing (Figure 4, lanes 1 and 3). Based upon this positive assessment of the presence of $(CA/GT)_n$ sequences, a library of Williams soybean DNA was screened for the presence of $(CA/GT)_n$ sequences.

Of approximately 10,000 clones examined, 36 produced a positive signal. Four of the 36 gave an obviously stronger signal than the remaining 32 and these clones were sequenced along with two of the remaining 32 which had given a weaker signal. These weakly hybridizing clones both contained a $(CA/GT)_4$ repeat. In the four clones producing stronger signals, $(CA/GT)_n$ repeats were also present and varied in length from n = 7 to n = 9 (Table 2). In addition, one sequence (45C5) also contained an $(AT)_4$ adjacent to the (CA) repeat and an $(ATT)_7$ 90 bp from the (CA) SSR.

Results of PCR amplification with primers de-

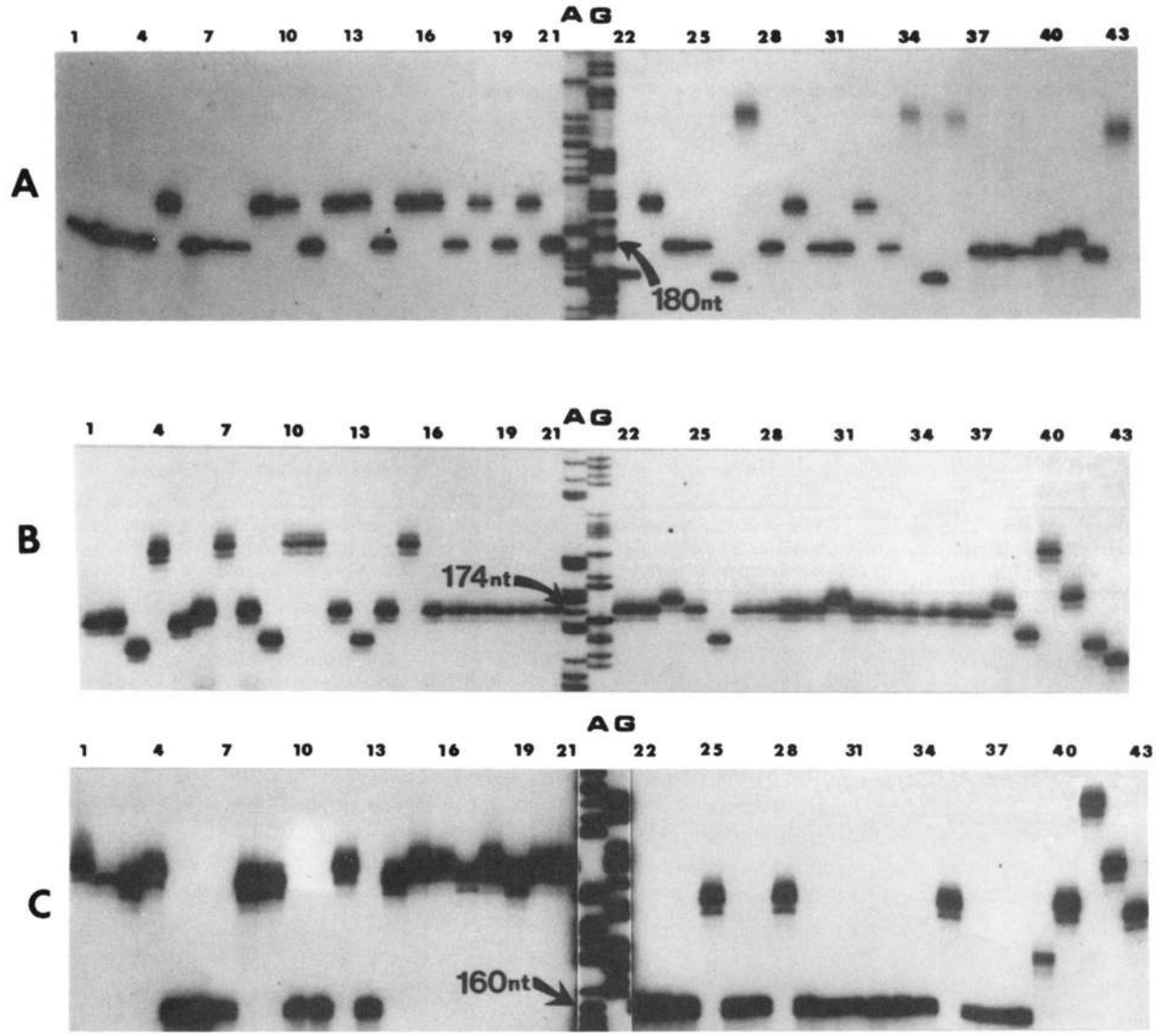

FIGURE 1.—Radiolabeled PCR products amplified from 43 soybean genotypes using primers flanking simple sequence repeats at three soybean loci, SOYHSP176$(AT)_n$ (A), SOYSC514$(AT)_n$ (B) and SOYPRP1$(ATT)_n$ (C), from GenBank. The genotypes in each lane are listed in Table 1. Fragments were separated on a standard DNA sequencing gel. The sequencing ladder between lanes 21 and 22 is the A and G lanes obtained from sequencing reactions using M13mp18 ssDNA.

rived from Bluescript clones containing $(CA/GT)_{>6}$: DNA of the cultivars Amsoy, Williams and Tokyo was used in an initial amplification using primers to selected Bluescript clones (Table 2). Multiple amplification products which were not polymorphic resulted in the case of clones 16C1 and 105H7. One product whose size was identical for the three soybean genotypes resulted from PCR using primers to 5A2. In the case of 45C5, a larger spectrum of soybean genotypes was examined. One genotype among a group of 19 produced a PCR product with a slightly higher molecular weight than the others. Because of the minimal level of SSR length polymorphism we did not investigate this locus further.

DISCUSSION

SSR length polymorphism in soybean: The degree of SSR length polymorphism at the SOYHSP176-$(AT)_n$, SOYSC514$(AT)_n$ and SOYPRP1$(ATT)_n$ loci was comparable to that found by WEBER and MAY (1989) at SSR loci in humans. They analyzed 10 loci and observed that the number of alleles ranged from 4 to 11 with an average of 6.8. At the three soybean loci examined here, an average of 7 alleles per locus was detected. WEBER and MAY (1989) calculated that the size difference between the largest and smallest allele at each locus ranged from 6 to 36 bp, with an average of about 16 bp. In soybean, the range in length appeared to be somewhat greater even if only the dinucleotide SSR loci SOYHSP176$(AT)_n$ and SOYSC514$(AT)_n$ are considered (Table 3). Whether our sampling of soybean genetic variation is comparable with the variation sampled by WEBER and MAY (1989) is difficult to determine. However, the degree of polymorphism at the three soybean loci investigated

TABLE 3

Number of soybean genotypes carrying each allele at three simple sequence repeat loci

Allele length	Locus and SSR sequence		
	SOYHSP176 $(AT)_n$	SOYSC514 $(AT)_n$	SOYPRP1 $(ATT)_n$
$Z-10$		1	
$Z-8$	3		
$Z-6$		5	
$Z-4$		1	
$Z-2$		1	
Z	22[a]	25[b]	20[c]
$Z+2$	1	3	
$Z+4$	1	1	
$Z+12$	12		1
$Z+14$		6	
$Z+21$			1
$Z+24$			4
$Z+27$			7
$Z+30$			9
$Z+42$	4		
$Z+51$			1
No. of alleles	6	8	7

The length (number of nucleotides) of the most common allele was arbitrarily assigned a value of Z. The number of nucleotides in alleles shorter or longer than Z are given by − and + values.

[a] $Z = 180$ nucleotides.

[b] $Z = 174$ nucleotides.

[c] $Z = 160$ nucleotides.

here appeared to be quite similar to polymorphic SSR loci in humans.

The 43 soybean genotypes included in the current study were selected to represent wide genetic diversity. If genotypes were put into classes based upon their allelic profiles, that is, those carrying identical alleles at each of the three loci would be placed in the same class, a total of 22 classes was required to accommodate the 43 genotypes. The *G. soja* genotypes each fell into unique classes. It was anticipated that these genotypes would be distinct from the *G. max* genotypes. Because the *G. soja* accessions were selected to represent diverse origins, it was also not surprising that they differed from each other. The cultivated soybeans fell into 17 allelic classes with up to 5 genotypes in one class. This group included the cultivars Mandarin, Lincoln, Clark, Williams and Douglas. Based upon their pedigrees, these genotypes are closely related. Similarly, the closely related cultivars S-100, Pickett and Lee were in a group by themselves. Thus, even the small amount of data collected at the three polymorphic SSR loci exhibiting length polymorphism allowed the grouping of genotypes into classes that would be anticipated based upon their species, origins, and pedigrees. With similar data from additional loci it should be possible to develop unique allelic profiles for establishing individual cultivar identity.

Simple sequence repeat DNA in soybean: The total length of the soybean sequences searched in GenBank was approximately 200 kbp. The presence of only four $(CA/GT)_n$ sequences with n ranging from only 5 to 8 was much less than anticipated based on human DNA sequence data. In the human genome various estimates indicate between 50,000 and 100,000 $(CA/GT)_n$ repeats with n ranging from 15 to 30 (BRAATEN *et al.* 1988; HAMADA and KAKUNAGA 1982; HAMADA, PETRINO and KAKUNAGA 1982; HAMADA *et al.* 1984; STALLINGS *et al.* 1991). Assuming equal spacing throughout the human genome, one such SSR would occur every 30–60 kbp. This was not the case in the soybean sequences contained in GenBank, as no $(CA/GT)_n$ with n greater than 8 was found.

The lack of such $(CA/GT)_n$ sequences in soybean was further supported by the results obtained in screening and sequencing clones in the library of Williams soybean DNA. About 3100 kbp of DNA was screened and 36 clones putatively containing $(CA/GT)_n$ sequences were identified. Our sequencing data suggest that in the 32 clones that gave weaker signals when probed with a ^{32}P-labeled poly(CA) probe, the length of $(CA/GT)_n$ was probably only 4 or 5 repeat units. By chance alone, one $(CA/GT)_4$ should occur every 65.5 kbp. Therefore, our finding of 32 such sequences in 3100 kbp was somewhat less than what would be anticipated. Of four clones that gave an obviously stronger signal with the poly(CA) probe, the maximum number of $(CA/GT)_n$ repeats was nine. If the frequency and length of $(CA/GT)_n$ SSRs in soybean were similar to that in human, we would have expected between 50 and 100 $(CA/GT)_n$ SSRs ranging in length from 15 to 30 repeat units. We found no evidence to suggest that such sequences exist in soybean.

The only data concerning frequency and length of dinucleotide repeats in a sizable sample of higher plant DNA is that of CONDIT and HUBBELL (1991). In two tropical tree species and *Zea mays* they estimated an average of one $(CA/GT)_n$ repeat every 103 kbp. This estimate was based on a library consisting of DNA fragments whose average size was assumed to be 256 bp. Their detection procedure used a radiolabeled probe and a low stringency similar to that employed in our screening of the Bluescript library of Williams DNA. If the 36 clones we detected all contained $(CA/GT)_n$ sequences, then one such sequence occurred every 86 kbp. This is similar to the result of CONDIT and HUBBELL (1991) and is not greatly different from estimates of $(CA/GT)_n$ sequences in humans. The more important question concerns the length of these SSRs. The longest $(CA/GT)_n$ sequence reported by CONDIT and HUBBELL (1991) had an n of 11. The longest such sequence we found had only 9 dinucleotide repeats. Together these experimental data fail to support the conclusion that long $(CA/GT)_n$ SSRs ($n > 14$) exist in plant species as they do in humans.

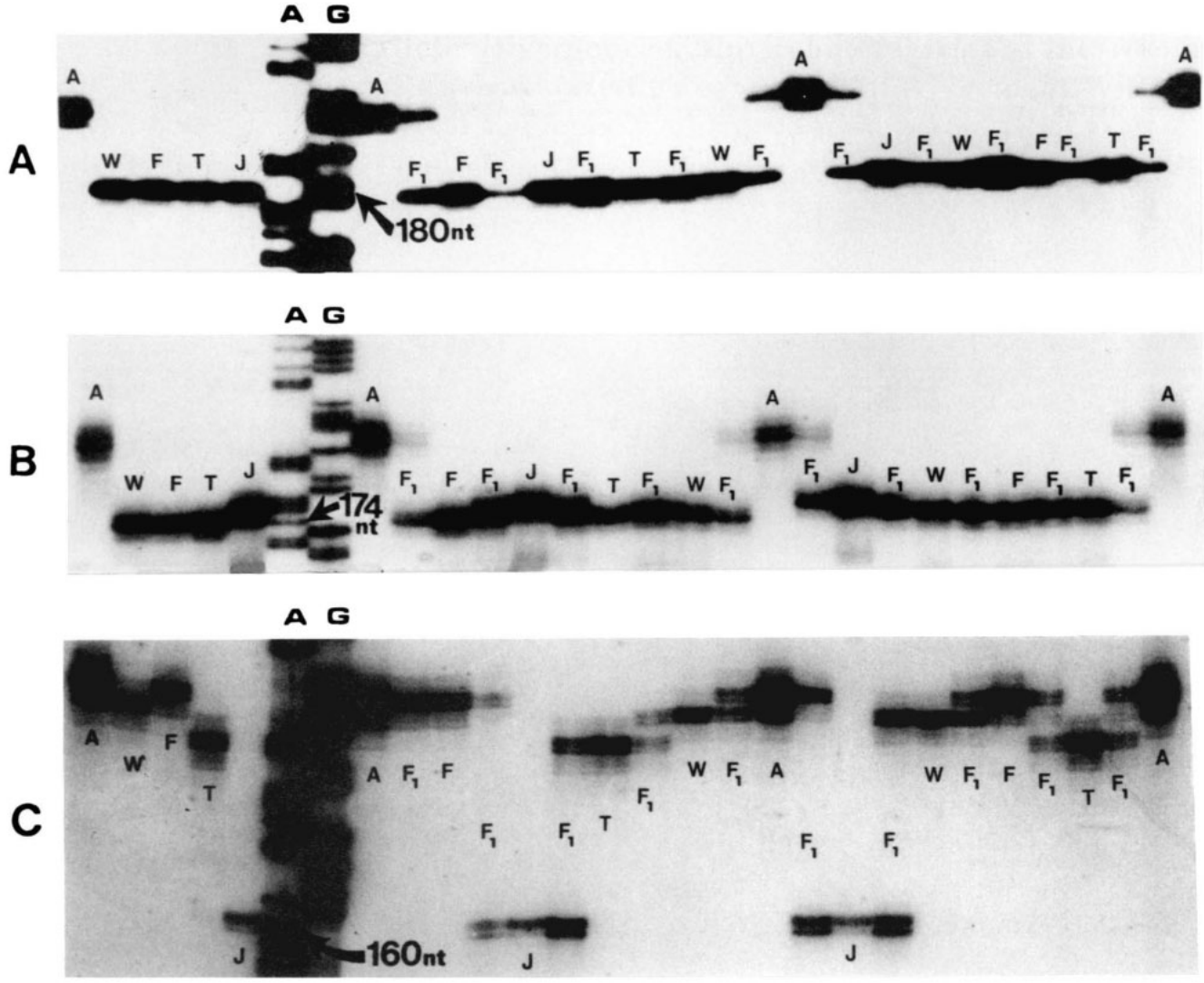

FIGURE 2.—Radiolabeled PCR products amplified from five selected soybean genotypes and the 10 F_1 hybrids from all possible crosses between them produced using primers flanking simple sequence repeats at three soybean loci SOYHSP176$(AT)_n$ (A), SOYSC514$(AT)_n$ (B) and SOYPRP1$(ATT)_n$ (C) from GenBank. Fragments were separated on a standard DNA sequencing gel. The sequencing ladder is the A and G lanes obtained from sequencing reactions using M13mp18 ssDNA. In the 21 lanes to the right of the sequencing ladder each F_1 is flanked by both of its parents (A = Amsoy, W = Williams, F = Fiskeby, T = Tokyo and J = Jackson soybean).

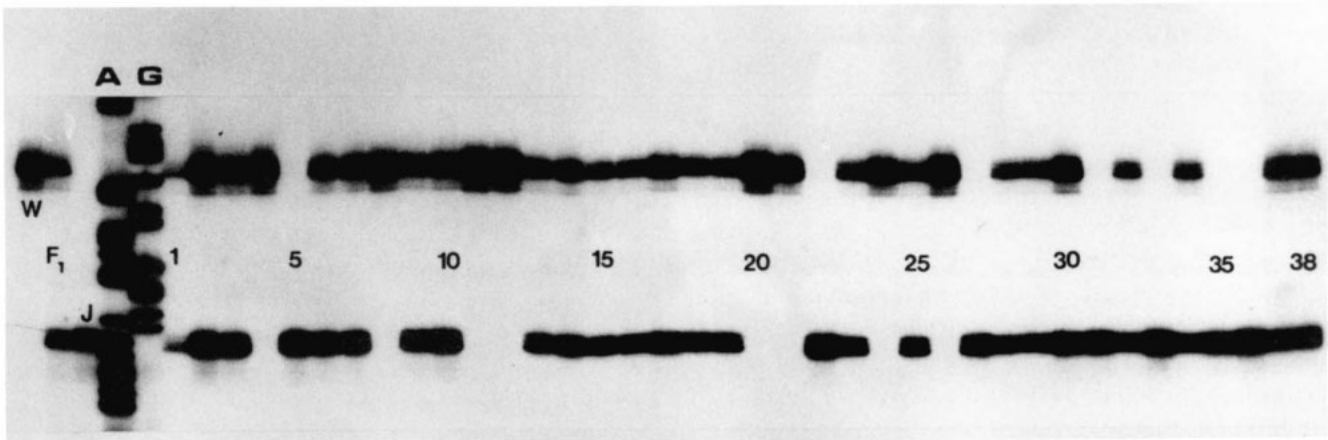

FIGURE 3.—Radiolabeled PCR products amplified from the DNA of the parents (J = Jackson, W = Williams), F_1, and 38 F_2 plants from a cross of Jackson × Williams soybean using primers to the simple sequence repeat locus SOYPRP1$(ATT)_n$. The sequencing ladder is the A and G lanes obtained from sequencing reactions using M13mp18 ssDNA.

TABLE 4

Segregation of simple sequence repeat alleles in the F_2 progeny of two soybean crosses

Locus	No. of F_2 plants	No. of F_2 plants in genotypic classes			χ^2	P
		P_1	Heterozygote	P_2		
Jackson (P_1) × Williams (P_2)						
SOYSC514(AT)$_n$	99	26	44	29	1.4	0.3–0.5
SOYPRP1(ATT)$_n$	98	23	53	22	0.7	0.7–0.9
Williams (P_1) × Amsoy (P_2)						
SOYHSP176(AT)$_n$	71	14	39	18	1.1	0.5–0.7
SOYSC514(AT)$_n$	71	14	32	25	4.1	0.1–0.2
SOYPRP1(ATT)$_n$	70	12	35	23	3.5	0.1–0.2

Progeny were classified as: homozygous for the allele from the maternal parent (P_1), heterozygous, or homozygous for the allele from the male parent (P_2), with an expected codominant segregation ratio of 1:2:1.

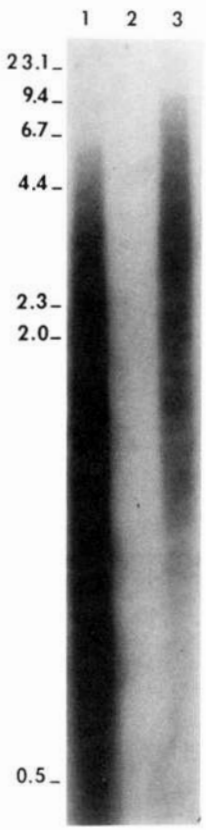

FIGURE 4.—Hybridization of a ^{32}P-labeled G(AC)$_{11}$ probe to *Pal*I digested human placental DNA (lane 1), *E. coli* strain B DNA (lane 2) and Williams soybean DNA (lane 3). After hybridization the membrane was washed five times including a final wash for 2 hr at 66° in 0.1 × SSC. Values in the left margin are in kilobases.

While the soybean sequences in GenBank had no (CA/GT)$_n$ sequences with n of 15 or greater, the five (AT/TA)$_n$ sequences with n ranging from 14 to 27 provided an excellent source of dinucleotide SSRs. The frequency of these sequences (1 per approximately 40 kbp) is similar to the frequency of (CA/GT)$_n$ SSRs of comparable length in humans.

CONCLUSIONS

Two findings emerge from the experiments reported here. First, the presence of polymorphic length SSRs in soybean demonstrates that this type of genetic maker is useful in a higher plant species. The average of 7 alleles present at each locus indicates that these should function as highly informative markers comparable to those in humans. The presence of SSR length polymorphism offers a complement to the RFLP and RAPD markers now commonly in use in many plant species. The rapidity of PCR combined with the informativeness of SSR length polymorphisms should make SSR markers valuable tools for molecular map development, genotype identification, and other uses. The highly polymorphic nature of SSR markers make them particularly advantageous in a species such as soybean in which RFLP has been somewhat difficult to detect.

The second result of interest relates to the lack of (CA/GT)$_n$ sequences in soybean comparable to that found in humans. While the frequency of such sequences may be similar to that found in human DNA, their length is apparently significantly less. Obviously, this fact has important implications on the use of (CA/GT)$_n$ SSRs as genetic markers in plants. WEBER (1990) demonstrated that the polymorphism information coefficient was near zero when n was 10 or fewer repeats, but reached approximately 0.8 with 24 or more repeats. Thus, it would appear that in soybean, and perhaps in other plants, the length of the (CA/GT)$_n$ sequences will not offer sufficient polymorphism to permit their use as genetic markers. However, the relative abundance of (AT/TA)$_n$ with n of 15 or greater in GenBank soybean sequences offers an al-

ternative to the use of $(CA/GT)_n$ SSRs as genetic markers. Trimeric repeats such as $(ATT)_n$ should also serve as a source of polymorphic SSRs although such sequences may not be as common in soybean as has been reported in the human genome (EDWARDS *et al.* 1991).

We thank PAMELA PRIGG ABUTALEB and EDWARD FICKUS for their technical assistance. The authors wish to thank DAVID NEAL, URI LAVI, LINCOLN MCBRIDE and JANET ZIEGLE for their careful reading of this manuscript. We gratefully acknowledge the U.S. Department of Agriculture, Agriculture Research Service Postdoctoral Research Associate Program for support of M.S.A.

LITERATURE CITED

BORTHWICH, H. A., and M. W. PARKER, 1952 Light in relation to flowering and vegetative development. pp. 801–810 in *Report of the 13th International Horticultural Congress of the Royal Horticultural Society*, London.

BOTSTEIN, D., R. L. WHITE, M. SKOLNICK and R. W. DAVIS, 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. **32:** 314–331.

BRAATEN, D. C., J. R. THOMAS, R. D. LITTLE, K. R. DICKSON, I. GOLDBERG, D. SCHLESSINGER, A. CICCODICOLA and M. D'URSO, 1988 Locations and contents of sequences that hybridize to poly(dgG-dT)·(dC-dA) in mammalian ribosomal DNAs and two X-linked genes. Nucleic Acids Res. **16:** 865–881.

CONDIT, R., and S. P. HUBBELL, 1991 Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. Genome **34:** 66–71.

EDWARDS, A., A. CIVITELLO, H. A. HAMMOND and C. T. CASKEY, 1991 DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am. J. Hum. Genet. **49:** 746–756.

HAMADA, H., and T. KAKUNAGA, 1982 Potential Z-DNA forming sequences are highly dispersed in the human genome. Nature **298:** 396–398.

HAMADA, H., M. G. PETRINO and T. KAKUNAGA, 1982 A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc. Natl. Acad. Sci. USA **79:** 6465–6469.

HAMADA, H., M. G. PETRINO, T. KAKUNAGA, M. SEIDMAN and B. D. STOLLAR, 1984 Characterization of genomic poly(dT-dG)·poly(dC-dA) sequences: structure, organization, and conformation. Mol. Cell. Biol. **4:** 2610–2621.

HONG, J. C., R. T. NAGAO and J. L. KEY, 1987 Characterization and sequence analysis of a developmentally regulated putative cell wall protein gene isolated from soybean. J. Biol. Chem. **262:** 8367–8376.

JACOB, H. J., K. LINDPAINTNER, S. E. LINCOLN, K. KUSUMI, R. K. BUNKER, YI-PEI MAO, D. GANTEN, V. J. DZAU and E. S. LANDER, 1991 Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. Cell **67:** 213–224.

JEFFREYS, A. J., V. WILSON and S. L. THEIN, 1985 Hypervariable "minisatellite" regions in human DNA. Nature **314:** 67–73.

JEFFREYS, A. J., V. WILSON, R. NEUMANN and J. KEYTE, 1988 Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. Nucleic Acids Res. **16:** 10953–10971.

KRAFT, R., J. TARDIFF, K. S. KRAUTER and L. A. LEINWAND, 1988 Using mini-prep plasmid DNA for sequencing double stranded templates with sequenase. Biotechniques **6:** 544–546.

LITT, M., and J. A. LUTY, 1989 A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. **44:** 397–401.

MULLIS, K., F. FALOONA, S. SCHARF, R. SAIKI, G. HORN and H. ERLICH, 1986 Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. Cold Spring Harbor Symp. Quant. Biol. **51:** 263–273.

NAGAO, R. T., E. CZARNECKA, W. B. GURLEY, F. SCHOEFFL and J. L. KEY, 1985 Genes for low-molecular weight heat shock proteins of soybeans: Sequence analysis of a multigene family. Mol. Cell. Biol. **5:** 3417–3428.

NAKAMURA, Y., M. LEPPERT, P. O'CONNELL, R. WOLFF, T. HOLM, M. CULVER, C. MARTIN, E. FUJIMOTO, M. HOFF, E. KUMLIN and R. WHITE, 1987 Variable number tandem repeat (VNTR) markers for human gene mapping. Science **235:** 1616–1622.

ROSTAS, K., E. KONDOROSI, B. HORVATH, A. SIMONCSITS and A. KONDOROSI, 1986 Conservation of extended promoter regions of nodulation genes in *Rhizobium* Proc. Natl. Acad. Sci. USA **83:** 1757–1761.

SAGHAI-MAROOF, M. A., K. M. SOLIMAN, R. A. JORGENSEN and R. W. ALLARD, 1984 Ribosomal DNA spacer length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA **81:** 8014–8019.

SAMBROOK, J., E. F. FRITSCH and T. MANIATIS, 1989 *Molecular Cloning: A Laboratory Manual*, Ed. 2. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

SHIBATA, D., T. KATO and K. TANAKA, 1991 Nucleotide sequence of a soybean lipoxygenase gene and the short intergenic region between an upstream lipoxygenase gene. Plant Mol. Biol. **16:** 353–359.

STALLINGS, R. L., A. F. FORD, D. NELSON, D. C. TORNEY, C. E. HILDEBRAND and R. K. MOYZIS, 1991 Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. Genomics **10:** 807–815.

TAUTZ, D., 1989 Hypervariability of simple sequences as a general source of polymorphic DNA markers. Nucleic Acids Res. **17:** 6463–6471.

TAUTZ, D., M. TRICK and G. A. DOVER, 1986 Cryptic simplicity in DNA is a major source of genetic variation. Nature **322:** 652–656.

WEBER, J. L., 1990 Informativeness of human $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ polymorphisms. Genomics **7:** 524–530.

WEBER, J. L., and P. E. MAY, 1989 Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. **44:** 388–396.

WILLIAMS, J. G. K., A. R. KUBELIK, K. J. LIVAK, J. A. RAFALSKI and S. V. TINGEY, 1990 DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. **18:** 6531–6535.

Communicating editor: S. D. TANKSLEY